SYNTHETIC ESTIMATES OF LITERACY PROFICIENCY FOR SMALL CENSUS AREAS

Stephen Reder[1]
Portland State University
P.O. Box 751
Portland, Oregon 97207
503-725-3999
reders@pdx.edu

Prepared for:

Division of Adult Education and Literacy
Office of Vocational and Adult Education
U.S. Department of Education[2]

Revised for Internet Publication
October 1997

INTRODUCTION

Adult literacy is increasingly seen as indispensable to the social and economic health of the United States. The *Goals 2000: Educate America* legislation (P.L.103-227) recognizes the importance of adult literacy as one of eight national education goals. Goal 6 states that "every adult will be literate and have the skills to compete in the global economy and participate in American democracy." The National Education Goals Panel, authorized by this legislation to monitor progress toward the goals, has adopted a specific set of adult literacy proficiency measures as the yardstick by which to judge progress toward meeting the adult literacy and lifelong learning goal (National Education Goals Panel, 1993a). These measures, developed by the Educational Testing Service (ETS), have been used in a number of state, national and international surveys of adult literacy over the past decade. In 1992, the measures were used to assess the literacy capabilities of the nation's adults age 16 and over. This National Adult Literacy Survey (NALS), carried out by ETS under contract to the National Center for Education Statistics, profiled the literacy abilities of the nation's adults on three scales: Prose, Document and Quantitative literacy (Kirsch, Jungeblut, Jenkins & Kolstad, 1993).

NALS surveyed a random sample of nearly 25,000 adults age 16 and over across the country.[3] Individuals were interviewed in their homes, providing rich background information about demographic characteristics, languages spoken, educational and occupational experiences, and their perceptions of and uses of literacy. The NALS also directly assessed respondents' abilities to perform everyday literacy tasks such as interpreting graphs and charts, extracting needed information from prose materials, completing forms, and so forth. Performance of these simulated tasks, which generally required constructed as opposed to multiple choice responses, was used to estimate individuals' Prose, Document and Quantitative proficiencies, each reported on a 0 to 500 scale. Five performance levels were designated on each scale: Level 1 (225 and under), Level 2 (226-275), Level 3 (276-325), Level 4 (326-375) and Level 5 (above 375). (Kirsch et al, 1993).

The National Education Goals Panel adopted these three proficiency scales - Prose, Document and Quantitative - as the indicators of progress toward meeting Goal 6. Both the mean proficiency of the adult population and the percentage performing at the two lowest levels are seen as useful indicators. Individual states have been utilizing these measures to monitor and report their own progress toward meeting Goal 6 (National Education Goals Panel, 1993b).

To obtain useful information about the literacy abilities and needs of their adult populations, a number of states contracted with ETS to conduct concurrent state adult literacy surveys (SALS) as part of the NALS.[4] A few other states have conducted related efforts.[5] To assist states that did not have SALS or SALS-like surveys, the Office of Vocational and Adult Education contracted for the development of techniques for

estimating adult literacy proficiencies from 1990 U.S. Census data (Reder, 1994b). Those methods provided reasonably accurate estimates of state-level literacy proficiencies.

Although these synthetic state-level estimates were useful for characterizing overall state needs and progress in relation to adult literacy goals, state and local programs have found themselves in need of more locally focused data as decision-making, priority-setting and allocation of limited programmatic resources are increasingly taking place at the state and local levels. The present work is thus the outgrowth of the increasing demand for information about adult literacy proficiencies and needs in more geographically focused areas. Techniques are developed and implemented in this paper which produce relatively accurate estimates of adult literacy proficiency at the level of individual counties, congressional districts, and cities, towns and places having at l east 10,000 inhabitants.

METHOD

Approach

The approach used here is similar to that used in the earlier synthetic estimation work of Reder (1994b). The previous work involved using regression models to predict *individual* NALS literacy proficiencies from individual background variables that are closely aligned with the 1990 U.S. Census long-form questions. These regression models were then applied to the 5% sample of Public Use Microdata Samples (PUMS) from the 1990 U.S. Census. The individual records in PUMS are sampled at random from within Public Use Microdata Areas (PUMAs), and so the literacy predictions that can be generated by applying the regression models to the PUMS records can be aggregated at the PUMA area. Unfortunately, the PUMA areas for which such synthetic literacy estimates can be generated are often not well aligned with the service areas or geographical units of interest to adult literacy programs.

The present approach utilizes a related but variant technique to produce synthetic estimates for a wider variety of Census areas. Rather than developing regression models that predict *individual* literacy proficiencies from individual PUMS records, the present approach develops statistical models which predict the literacy proficiencies of *populations* of individuals from their *aggregate* characteristics (i.e., from their collective profile in terms of demographics, educational experiences, occupations, etc.). Such models are then applied to published summary tabulations of long-form Census data for a variety of Census areas, generating estimates of literacy proficiencies for those areas.

To develop such models, individual records in the NALS data set are first aggregated into counties (the only local geographic identifiers available in the NALS records) whose aggregate literacy characteristics can be modeled in relation to background variables that can be closely aligned with 1990 U.S. Census long-form variables. The regression models can then utilized to predict literacy proficiencies for other aggregates in the summary Census tabulations.

Details of these methods, of their validation, and of the results they produce are described below.

Data Sources

Two data sets were used to develop the synthetic estimates, one from the National Adult Literacy Survey (NALS), and one from the 1990 U.S. Census. A data tape for the NALS, provided by the Educational Testing Service, was utilized to develop regression models for predicting county-level NALS literacy proficiencies from aggregated responses to NALS background questionnaires. These regression models were then applied to summary tables of corresponding variables in the 1990 U.S. Census (long form) to generate predicted values, standard errors and confidence intervals for literacy proficiencies at the county, town/city and congressional district levels. The Census data used were extracted from the CD-ROM versions of Summary Tape File 3C for counties, cities and towns and Summary Tape File 3D for the congressional districts of the 103rd Congress.

Variable Alignment

The valid application of regression models predicting assessed NALS literacy proficiencies to predicting literacy proficiencies from Census data requires the use of a set of common predictor variables that are closely aligned across the two data sets. By design, the NALS included numerous variables common to the long-form of the 1990 Census. The information the NALS background and Census long-form questionnaires collect in common describe such demographic characteristics as age, gender, place of birth, and educational attainment. Each questionnaire further collected information about labor force participation, employment and occupational status, income from various sources, languages spoken in the home (and ratings of oral English proficiency if other languages are spoken), marital status, household composition, and so forth.

Despite this rich potential overlap of information between NALS and long-form Census, there are several factors that limit the variables that can be closely aligned between the two data sets. Some information common to the two data sets cannot be used because the pertinent questions were not asked in a parallel fashion or recorded in terms of sufficiently similar response alternatives across the two data sets. Marital status, for example, cannot be used as a common predictor for this reason. Household poverty status, as another example, is not reported comparably in the two studies. Household-level as opposed to individual-level variables are generally difficult to align because of definitional and procedural differences between the Census and NALS.

Some variables could be made parallel across the two data sets by recoding them according to a common scale or set of response alternatives. *Age*, for example, is recorded as a continuous variable in NALS, and can thus be categorized into subranges that match the age categories in the summary Census tables. Another example is provided by the *recent immigrant* variable, which was recoded so that responses on both NALS and

long-from Census questionnaires could be aligned; a person not born in the United States was defined as being a recent immigrant if he or she had immigrated to the U.S. within a 5 year period preceding the NALS interview or 1990 Census-taking; the 5 year cut off point was one of a limited number of alignment points between the alternative response categories in the 1990 Census and NALS. Educational attainment, as a third example, was recoded into a set of discrete response categories that could be aligned. Some distinctions made in one questionnaire were not made in the other. For example, distinctions among advanced degrees (e.g., master's level versus doctoral level) were made in the Census but not in the NALS, whereas distinctions among small numbers of years of education are made in the NALS but not in the Census. The GED is distinguished from a high school diploma in the NALS but not in the 1990 Census. A set of six categories of educational attainment was constructed into which all responses on both NALS and Census could be unambiguously and uniquely mapped.

There are other limitations on aligning NALS and STF variables. Because the NALS includes only individuals age 16 and above, for example, some Census variables could not be closely aligned because they are tabulated in the STF files only for a different age range. For example, place-of-birth data in the STF files is tabulated for all persons regardless of age, whereas the same information in NALS is available only for persons 16 and above. This could potentially bias the alignment of this variable across geographical aggregates (e.g., 89.4 % of the NALS population - age 16 and above - were born in the United States, compared to 90.7 % of the cradle-to-grave Census population). Another subtle population difference between NALS and Census is that for many variables, STF tabulations include military, institutionalized, and “group quarters” individuals whereas NALS includes only household residents and not these other subpopulations. Other relatively small population differences are differences in whether college students living in dormitories are included and whether adjustments have been made for apparent undercount in the 1990 U.S. Census (Census of Population and Housing, 1992; Reder, 1994b).

Despite these and other relatively minor limitations identified in Table 1, the overall alignment of the two data sets proved satisfactory as evidenced by the modeling and validation studies presented below. Details of the common variables, their coding and their alignments across the two studies are presented in Table 1. Notice that these aligned model variables are organized as sets of proportions that sum to one; the variables that are grouped together in this way are boxed together by heavier horizontal lines in the table. For example, there are seven educational attainment variables, each measured as a proportion of the population that has a certain level of educational attainment (less than high school, some high school, ..., graduate school). The variables listed in each set are non-overlapping and their corresponding proportions always sum to one.

TABLE 1 - ALIGNED MODEL VARIABLES

PREDICTOR	NALS VARIABLE	STF3 TABLE(S)	COMMON CODING OVER POPULATION AGGREGATES
Educ less than high school	BLB0101	P57	Proportion of persons age 18 and above with less than high school education
Educ-some high school	BLB0101	P57	Proportion of persons age 18 and above with some high school education
Educ-high school diploma/GED	BLB0101	P57	Proportion of persons age 18 and above with a high school diploma, GED or equivalent
Educ-some college	BLB0101	P57	Proportion of persons age 18 and above with some college education (no degree)
Educ-2 year college degree	BLB0101	P57	Proportion of persons age 18 and above with a 2 year college degree
Educ-4 year college degree	BLB0101	P57	Proportion of persons age 18 and above with a 4 year college degree
Educ-graduate school	BLB0101	P57	Proportion of persons age 18 and above with graduate/professional school education
White	BNF0901	P14A,B	Proportion of persons age 16 and above identifying race as White
Black	BNF0901	P14C,D	Proportion of persons age 16 and above identifyng race as Black
Native American	BNF0901	P14E,F	Proportion of persons age 16 and above identifying race Native American
Asian/Pacific Islander	BNF0901	P14G,H	Proportion of persons age 16 and above identifying race as Asian or Pacific Islander
Other race	BNF0901	P14I, J	Proportion of persons age 16 and above identifying race as “other”
Age 16-24	DAGE	P13	Proportion of persons age 16 and above of age 16-24
Age 25-34	DAGE	P13	Proportion of persons age 16 and above of age 25-34
Age 35-44	DAGE	P13	Proportion of persons age 16 and above of age 35-44
Age 45-54	DAGE	P13	Proportion of persons age 16 and above of age 45-54
Age 55-64	DAGE	P13	Proportion of persons age 16 and above of age 55-64
Age 65 & above	DAGE	P13	Proportion of persons age 16 and above of age 65 & above
Hispanic	BG10701	P15A,B	Proportion of persons age 16 and above of Hispanic origin
Not Hispanic	BG10701	P15A,B	Proportion of persons age 16 and above not of Hispanic origin
Work disability	BLB1301	P66	Prop. of civilian noninstitutionalized pop. age 16 & up with a work disability
No work disability	BLB1301	P66	Prop. of civilian noninstitutionalized pop. age 16 & up without a work disability
Speaks English very well	DLANGBS BLA1501	P28	Proportion of persons age 18 and above who spoke language other than English before starting school who now speak English “very well”
Speaks English well	“	P28	“ “ “ ... who now speak English “well”
Speaks English not well/not at all	“	P28	“ “ “ ... who now speak English “not well” or “not at all”

TABLE 1 - ALIGNED MODEL VARIABLES (Continued)

PREDICTOR	NALS VARIABLE	STF3 TABLE(S)	COMMON CODING OVER POPULATION AGGREGATES
Recent immigrant	BNA0201	P36	Proportion of persons who immigrated to United States within preceding 5 years
Not recent immigrant	BNA0201	P36	Proportion of persons who did not immigrate to United States within preceding 5 years
U.S.-born	BLA0101	P42	Proportion of persons born in the United States
Not U.S.-born	BLA0101	P42	Proportion of persons born outside of in the United States
Did not work previous year	DWKSWRK	P76	Proportion of persons age 16 and above who did not work in previous year
Worked 1-13 weeks previous year	DWKSWRK	P76	Proportion of persons age 16 and above who worked 1-13 weeks during previous year
Worked 14-26 weeks previous year	DWKSWRK	P76	Proportion of persons age 16 and above who worked 14-26 weeks during previous year
Worked 27-39 weeks previous year	DWKSWRK	P76	Proportion of persons age 16 and above who worked 27-39 weeks during previous year
Worked 40-52 weeks previous year	DWKSWRK	P76	Proportion of persons age 16 and above who worked 40-52 weeks during previous year
Laborer	BLD1001	P78	Proportion of employed persons 16 and above in occupational class
Service	BLD1001	P78	Proportion of employed persons 16 and above in occupational class
Sales/administrative support	BLD1001	P78	Proportion of employed persons 16 and above in occupational class
Professional/technical/managerial	BLD1001	P78	Proportion of employed persons 16 and above in occupational class
Not in labor force	BLD01xx BD03901	P70	Proportion of persons not currently in labor force
Unemployed	" "	P70	Proportion of persons not currently working and looking for work
Employed	" "	P70	Proportion of persons currently employed (part-time or full-time)
Family income 0 - $4,999	BNF0701,02	P107	Proportion of families with previous year's income in specified range
Family income $ 5,000 - $9,999	BNF0701,02	P107	Proportion of families with previous year's income in specified range
Family income $10,000 - $14,999	BNF0701,02	P107	Proportion of families with previous year's income in specified range
Family income $15,000 - $19,999	BNF0701,02	P107	Proportion of families with previous year's income in specified range
Family income $20,000 - $29,999	BNF0701,02	P107	Proportion of families with previous year's income in specified range
Family income $30,000 - $39,999	BNF0701,02	P107	Proportion of families with previous year's income in specified range
Family income $40,000 - $49,999	BNF0701,02	P107	Proportion of families with previous year's income in specified range
Family income $50,000 - $74,999	BNF0701,02	P107	Proportion of families with previous year's income in specified range
Family income $75,000 & above	BNF0701,02	P107	Proportion of families with previous year's income in specified range

TABLE 1 - ALIGNED MODEL VARIABLES (Continued)

PREDICTOR	NALS VARIABLE	STF3 TABLE(S)	COMMON CODING OVER POPULATION AGGREGATES
Female	BG12401	P14A .. J	Proportion of persons age 16 and above who are female
Male	BG12401	P14A .. J	Proportion of persons age 16 and above who are male
Household size = 1	HHSIZE	P16	Proportion of households of given size
Household size = 2	HHSIZE	P16	Proportion of households of given size
Household size = 3	HHSIZE	P16	Proportion of households of given size
Household size = 4	HHSIZE	P16	Proportion of households of given size
Household size = 5	HHSIZE	P16	Proportion of households of given size
Household size = 6	HHSIZE	P16	Proportion of households of given size
Household size >= 7	HHSIZE	P16	Proportion of households of given size
Northeast	CENREG	REG	1 if in Northeast Census region else 0
Midwest	CENREG	REG	1 if in Midwest Census region else 0
South	CENREG	REG	1 if in South Census region else 0
West	CENREG	REG	1 if in West Census region else 0

County Aggregation

Standard federal state and county identifiers are provided in the NALS data set along with population sampling weights for each person in the survey sample. The sampling design of NALS sampled persons in households at geographically random points within a selected hierarchy of geographical strata (Kirsch et al, forthcoming).[6] For analytical purposes, the 24,944 NALS household survey respondents were aggregated into 417 unique counties; the number of survey respondents per county ranged between 3 and 902.

NALS data were aggregated over these 417 counties, including the predictor variables described in Table 1 and the mean values of the dependent variables of interest for this study: the mean combined NALS literacy proficiency; the proportion of individuals having combined literacy proficiency at Level 1 (i.e., 225 and below); and the proportion with combined proficiency at either Level 1 or 2 (i.e., 275 and below).[7] NALS case weights were used in calculating all aggregated values. This aggregated data file, now with 417 cases in it, one per county, was the analytical data set for the regression modeling described below.

Regression Modeling

Multiple linear regression techniques were used to predict the mean literacy proficiency for the county aggregates, the proportion of county scores at Level 1, and the proportion of county scores at Level 1 or 2. Separate regression models were developed for each of these dependent variables. Preliminary analyses indicated that better fitting and more robust regression models were obtained when county aggregates based on relatively small subsamples of respondents were excluded from the analyses. This should not be particularly surprising, since there is much more variability in the mean values of both independent an dependent variables aggregated over small subsamples. Analysis of regression residuals indicated that a reasonable threshold was 50 cases or more per county. Therefore, the models were developed and fitted to counties having 50 or more respondents in the survey. Of the 417 counties in the aggregate file, 178 met this criterion, whereas 239 had fewer than 50 cases and were excluded from the modeling process.

Weighted least squares (WLS) regression models yielded considerably better fits than ordinary least squares (OLS) models. Weighted least squares techniques are appropriate in cases where the dependent variable is heteroscedastic, i.e., does not have uniform variance at each point. Because the aggregates -- which were the units of analysis for these models -- were themselves based on varying numbers of observations, it seemed reasonable that the variance of the dependent variable being predicted would vary with the number of cases upon which it is based. Reasonable approximations to these variances would be proportional to 1/N for the mean literacy (where N is the number of cases in the given county subsample) and to p(1-p)/N for the fraction of individuals with scores below

a certain value, where p is the population proportion, estimated by the observed fraction in the sample of size N. If these formulas are reasonable approximations (up to a multiplicative constant) of the variances of the dependent variables, then the appropriate WLS weights should be inversely proportional to the variances, i.e., a weight proportional to N for the mean literacy equation and to p(1-p)/N for the fraction of cases below some threshold literacy value, where N is the county subsample size and p is the sample proportion of cases below the target literacy value.

Using these regression weights, highly predictive equations for the dependent variables were identified using common WLS regression techniques.[8] A number of transformations were applied to the dependent variables that were proportions, i.e., the proportion at Level 1 and the proportion at Level 1 or 2. Logit, probit, arcsin and square root transformations were applied to these dependent variables in an attempt to normalize their distributions and improve the fit of the regression models. But the best fitting models for these dependent variables turned out to be ones which directly predicted the simple proportions rather than some transformation of the proportions.

RESULTS

The variables appearing in the final (i.e., best-fitting) WLS regression equations are indicated in Table 2. Significant predictors are marked with an "x" in Table 2 in the column(s) corresponding to the equation(s) in which they play a statistically significant role. For example, each variable representing a different level of educational attainment is a significant predictor of mean proficiency. Notice that within each set of related variables, one (e.g., educ less than high school) is preceded by an "*" and is followed by a shaded row; this indicates that the variable was not included in the regressions, since it is a perfect linear combination of the others in the set (variables in a set always sum to 1).

Some variables listed in Table 1 -- age, gender, family income, household size, U.S. birth place -- do not appear in Table 2. That is because those variables are not statistically significant predictors of any of the three dependent variables. The fact that these variables do not appear predictive of the *aggregate* literacy data does *not* necessarily indicate mean they are not important predictors of *individual* literacy. For example, although age is a strong predictor of individual literacy (Kirsch et al, 1993; Reder, 1994b), it does not predict differences here among literacy scores at the county level. Apparently, existing differences in the age distribution of county populations are not strongly associated with differences among those counties in adult literacy.

TABLE 2 - Significant Predictor Variables in the Regression Models

PREDICTOR	MEAN PROFICIENCY	% AT LEVEL 1	% AT LEVEL 1 OR 2
*Educ less than high school			
Educ-some high school	x		x
Educ-high school diploma/GED	x	x	x
Educ-some college	x	x	x
Educ-2 year college degree	x		x
Educ-4 year college degree	x		x
Educ-graduate school	x	x	x
*White			
Black	x	x	x
Native American			
Asian/Pacific Islander			
Other race			
Work disability	x	x	x
*No work disability			
*Speaks English very well			
Speaks English well	x	x	x
Speaks English not well/not at all	x	x	
Recent immigrant	x		x
*Not recent immigrant			
*Did not work previous year			
Worked 1-13 weeks previous year	x		
Worked 14-26 weeks previous year			
Worked 27-39 weeks previous year			
Worked 40-52 weeks previous year			
*Laborer			
Service			
Sales/administrative support	x		
Professional/technical/managerial			
*Not in labor force			
Unemployed		x	
Employed	x	x	x
Northeast			x
Midwest	x		
South			x
*West			

The WLS regression models using these variables fit the county-level data extremely well, as shown in Table 3. For each equation the multiple R, adjusted R^2, and degrees of freedom (for the regression and residuals) are shown. With R values over .9 for each of the equations, we see that these regression models account for 81 to 91 % of the variance among counties in the literacy measures, by all accounts an excellent fit. The bottom row of the table displays the *maximum* value assumed by Cook's Distance over the 178 points

being fit; Cook's D is an indicator of how influential a given data point is on the regression equation, that is, how much the fit of the equation is influenced by a particular value. The small maximum values shown for Cook's D in the table (Cook's D is not bounded above by 1) is further evidence of a good-fitting model (Cook, 1977).

TABLE 3 - Summary of Fit of Regression Models

	Equation for MEAN PROFICIENCY	Equation for PROPORTION AT LEVEL 1	Equation for PROPORTION AT LEVEL 1 OR 2
Multiple R	.958	.904	.946
Adjusted R^2	.911	.808	.886
Degrees of Freedom	15 & 162	9 & 168	13 & 164
Maximum Cook's D	.099	.263	.078

Table 4 displays the unstandardized regression coefficients for the three equations. The complete regression equation is specified in each column, including the constant term displayed in the bottom row. Numerical coefficients shown in the table occur where the "x"s appeared previously in Table 2. Each is statistically significant (from zero) at the .05 level or better; blank cells in the table indicate that the corresponding coefficient is not statistically different from zero. As noted above, other variables considered in the modeling process that do not appear in the table were not significant predictors of any of the three dependent variables. Notice that negative signs on the coefficients in the mean literacy proficiency equation are associated with lower levels of average literacy, whereas negative coefficients in the other two equations are associated with higher levels of literacy (i.e., with smaller proportions of adults scoring at the lower levels of literacy). The corresponding coefficients for standardized independent variables (βs) are listed in Appendix A.

TABLE 4 - Unstandardized Coefficients for Regression Equations

PREDICTOR	MEAN PROFICIENCY	PROPORTION AT LEVEL 1	PROPORTION AT LEVEL 1 OR 2
*Educ less than high school			
Educ-some high school	79.61		-.382
Educ-high school diploma/GED	104.19	-.226	-.632
Educ-some college	123.99	-.292	-.787
Educ-2 year college degree	135.50		-1.062
Educ-4 year college degree	140.13		-.798
Educ-graduate school	181.57	-.398	-1.268
*White			
Black	-48.63	.330	.335
Native American			
Asian/Pacific Islander			
Other race			
Work disability	-25.36	.297	.267
*No work disability			
*Speaks English very well			
Speaks English well	-65.46	.414	.596
Speaks English not well/not at all	-60.95	.710	
Recent immigrant	-52.60		.487
*Not recent immigrant			
*Did not work previous year			
Worked 1-13 weeks previous year	71.15		
Worked 14-26 weeks previous year			
Worked 27-39 weeks previous year			
Worked 40-52 weeks previous year			
*Laborer			
Service			
Sales/administrative support	17.71	-.228	
Professional/technical/managerial		-.142	
*Not in labor force			
Unemployed		-.222	
Employed	32.45	-.288	-.295
Northeast			.028
Midwest	3.75		
South			.026
*West			
CONSTANT	149.13	.431	1.183

Figures 1, 2 and 3 display the relationships between the predicted and observed values for the three dependent variables. The strong correlation between observed and predicted values is evident in each of these scatterplots.

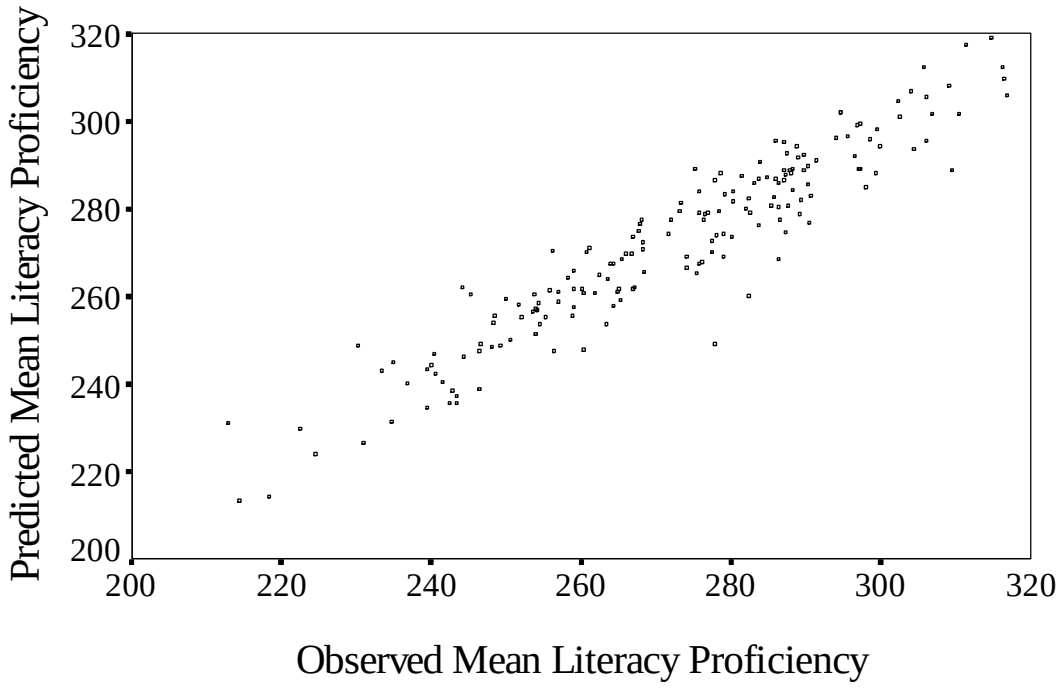

Figure 1. Scatterplot of predicted versus observed mean literacy proficiency for county aggregates.

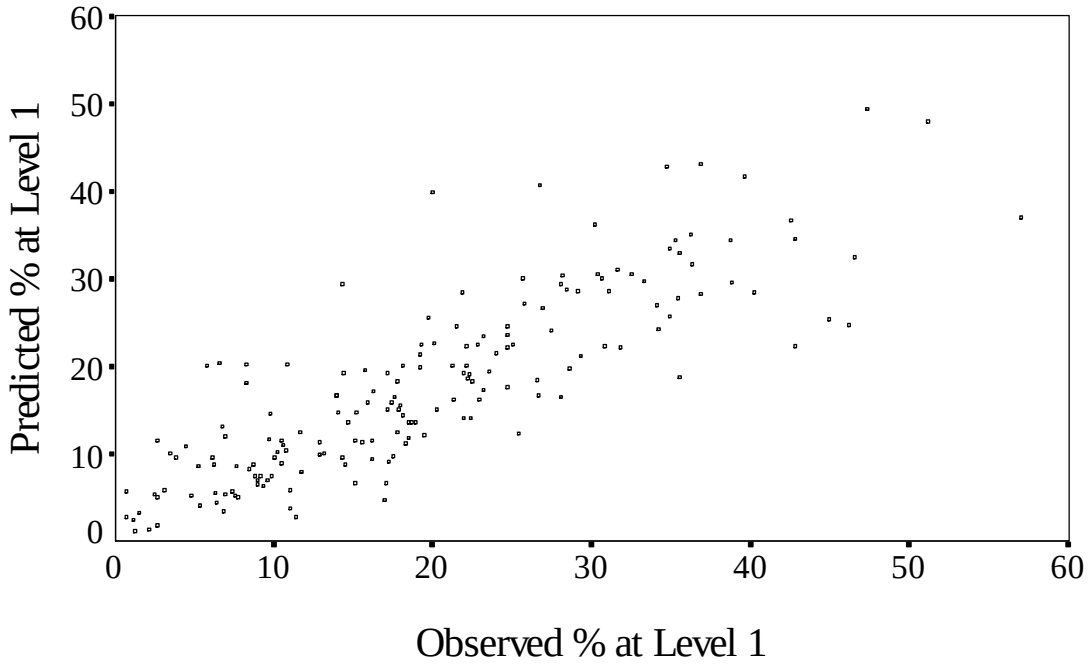

Figure 2. Scatterplot of predicted versus observed percent of adults in counties having combined literacy proficiency in Level 1.

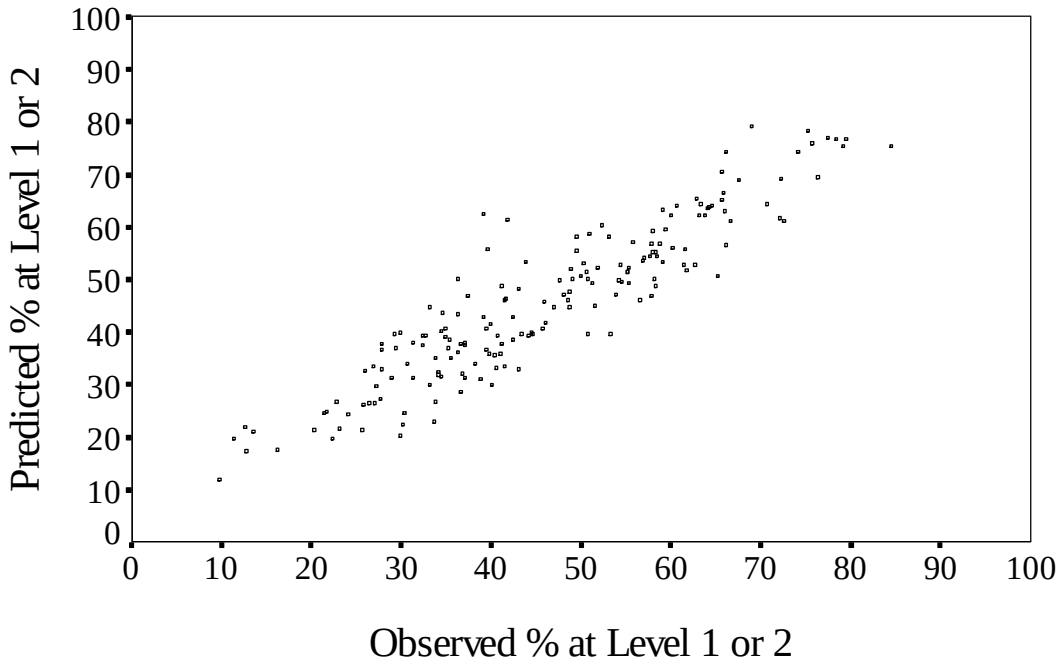

Figure 3. Scatterplot of predicted versus observed percent of adults in counties having combined literacy proficiency in Level 1 or 2.

Further information about the goodness of fit of these models is provided by analysis of the residuals of each equation. Figures 4, 5 and 6 exhibit scatterplots for the weighted residual by weighted predicted values for each county, one figure per dependent variable. The overall "shotgun blast" appearance of these scatterplots is additional evidence of how well the equations fit the county-level data.

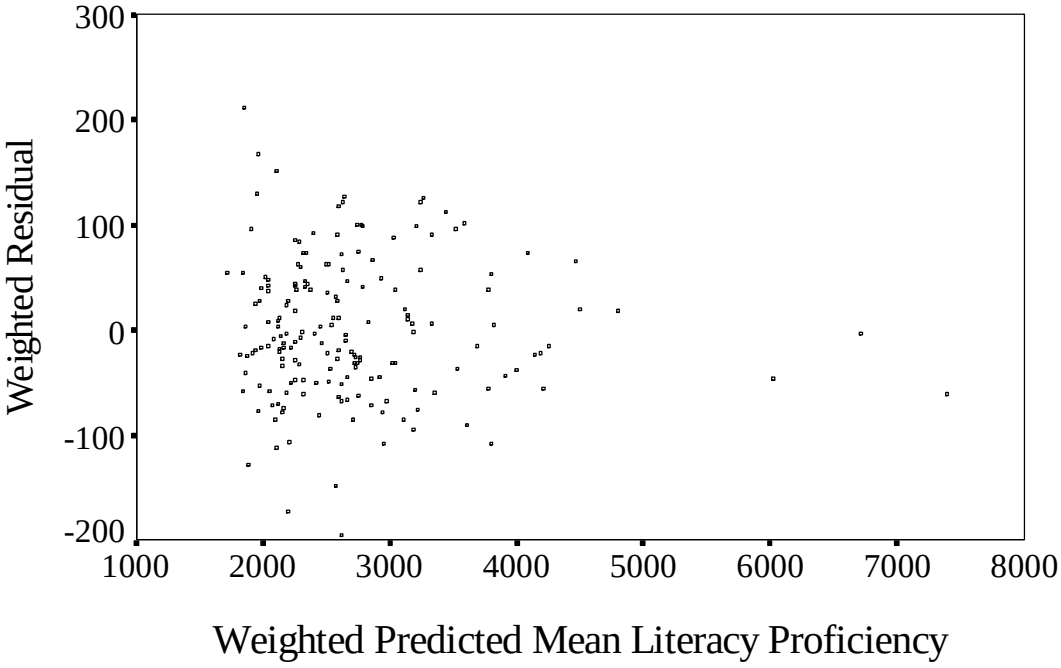

Figure 4. Scatterplot of weighted residuals versus weighted predicted values for mean literacy proficiency of counties.

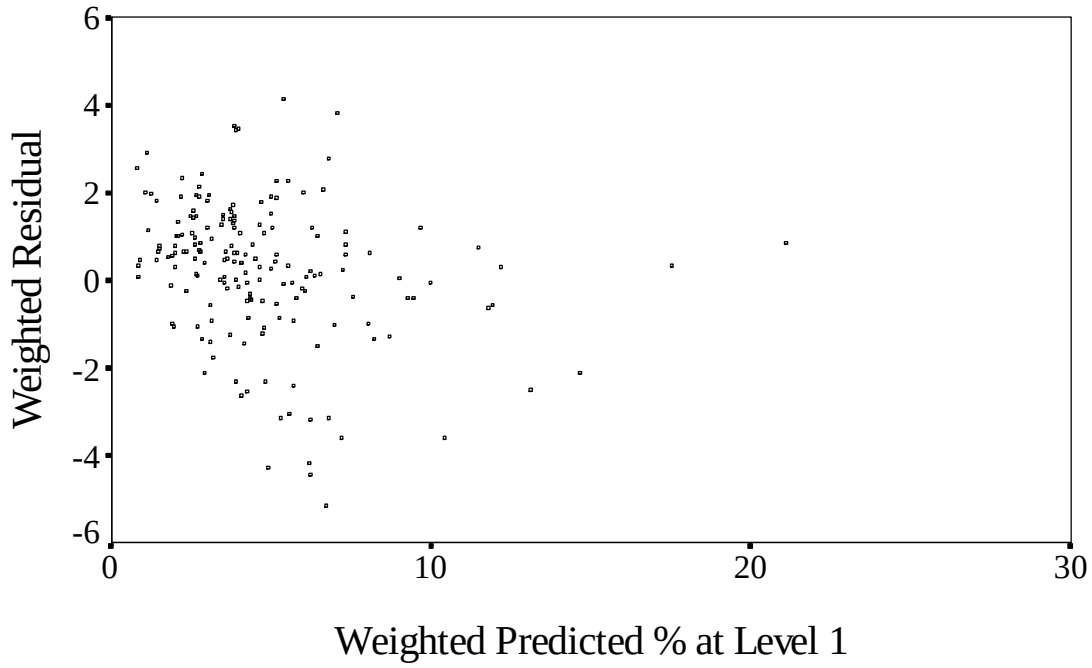

Figure 5. Scatterplot of weighted residuals versus weighted predicted percent of adults in counties having combined literacy proficiency in Level 1.

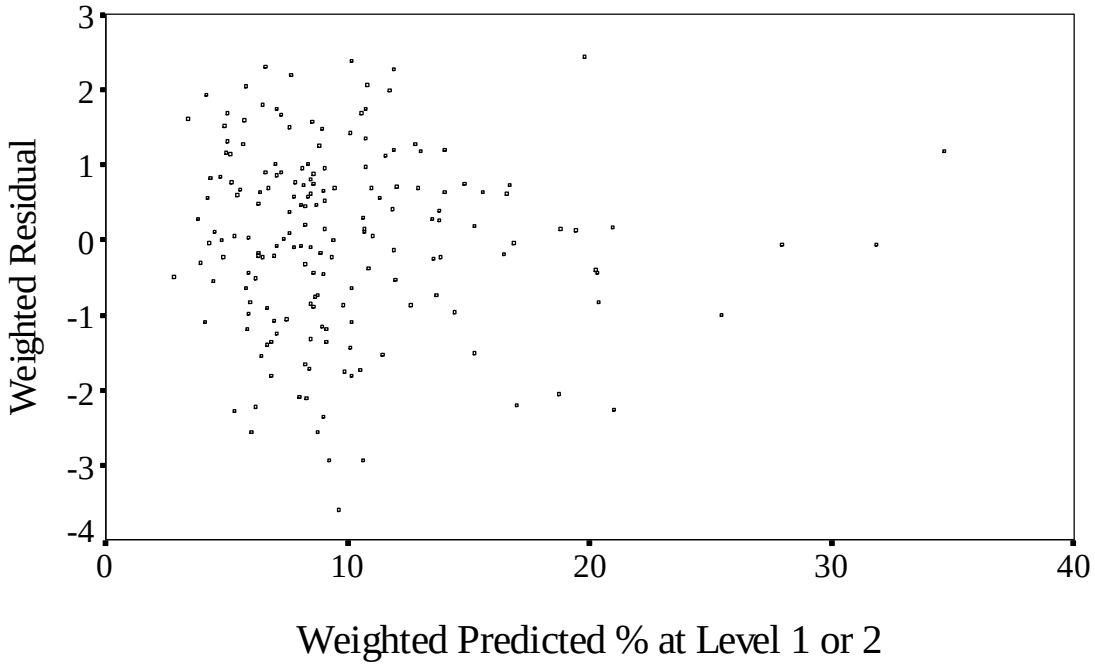

Figure 6. Scatterplot of weighted residuals versus weighted predicted percent of adults in counties having combined literacy proficiency in Level 1 or 2.

Generating Predictions for Small Census Areas

The regression equations exhibited in Table 4 were applied to 1990 Census Summary Tape File 3 data (recoded as specified in Table 1) to generate literacy predictions for Census areas. Predictions were generated for the entire population (STF Geocode 00) age 16 and above within a Census-defined county (STF Summary Level 050), county subdivision (Summary Levels 061 and 062), cities, towns and places of 10,000 or more inhabitants (STF Summary Levels 161 and 170), and congressional districts of the 103rd Congress (STF Summary Level 501). For purposes of keeping standard errors acceptably low among the Census variables used as predictors, estimates were generated only for those counties, cities, towns or places having at least 5,000 inhabitants age 16 and above and a realized sample of at least 500 for the long-form of the 1990 Census. Because cities, towns and places tabulated in STF3 have a minimum of 10,000 inhabitants, all 3,154 such units met the screening criteria. Of the 4,625 counties and county subdivisions in STF3, 4,026 passed the population and sample size criteria.[9]

In each area, three measures of adult literacy were estimated for the population age 16 and above: the mean combined NALS literacy proficiency; the percentage of persons with literacy proficiencies at Level 1; and the percentage of persons with literacy proficiencies at Levels 1 or 2. Each estimate generated was accompanied by a standard error and a 95% confidence interval for the individual prediction.[10] The confidence interval takes into account not only the inherent inaccuracy of the regression model's predictions, but also

the similarity, in terms of the predictor variables, of the given area to the NALS county aggregates on which the regression models were "trained"; the regression model tends to be less accurate for areas that are less similar to the NALS aggregates in terms of demographic and other predictive characteristics. Summary statistics for the standard errors and confidence intervals for each type of geographical unit are tabled in Appendix B.

Because of the large number of Census units for which these literacy estimates have been generated, they are being disseminated as electronic databases. Database files have been developed that can be viewed and printed with both personal computer software and standard Internet browsers. This software allow users to conveniently display and/or print out estimates, standard errors, and confidence intervals for the three literacy measures for Census areas, along with the local values of the predictor variables used by the equations (i.e., those listed in Table 2 or 4). The software allows users to examine the estimated literacy measures for selected states, congressional districts, counties, county subdivisions, cities, towns and places as defined by the Census STF3 geography. This software is available at several Internet locations.

Validation through SALS Comparisons

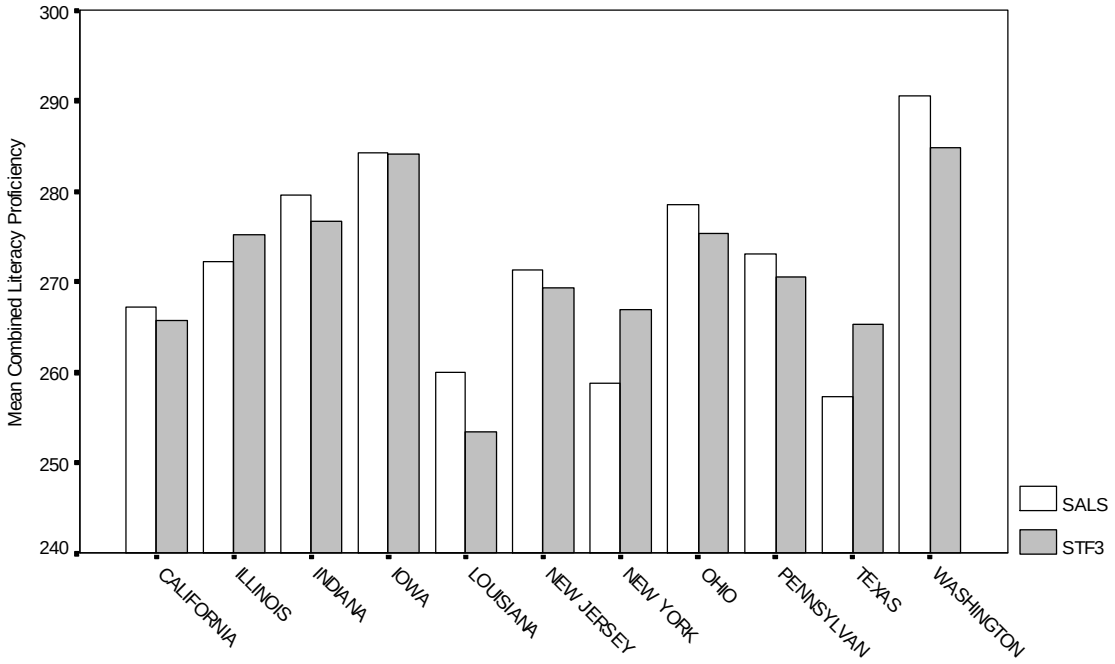

Figure 7. Comparison of synthetic estimates derived from Census STF3 data and State Adult Literacy Survey (SALS) estimates of statewide mean literacy proficiency.

The same procedures described above for generating literacy estimates for congressional districts, counties, cities, towns and places were also applied to state-level data in the STF3 files (Summary Level 040, Geocode 00). The statewide estimates can be compared

with corresponding statewide estimates made by the State Adult Literacy Survey (SALS) for those eleven states that contracted for concurrent state-valid surveys as part of the NALS.[11] Results of this comparison are displayed in Figure 7 for mean literacy proficiency, Figure 8 for percent at Level 1, and Figure 9 for percent at Level 1 or 2.

As can be seen from the figures, the regression model developed at the county level appears to fit the state level data for the SALS states reasonably well. Most of the state-level discrepancies are within the 95% confidence interval estimated by the models.[12]

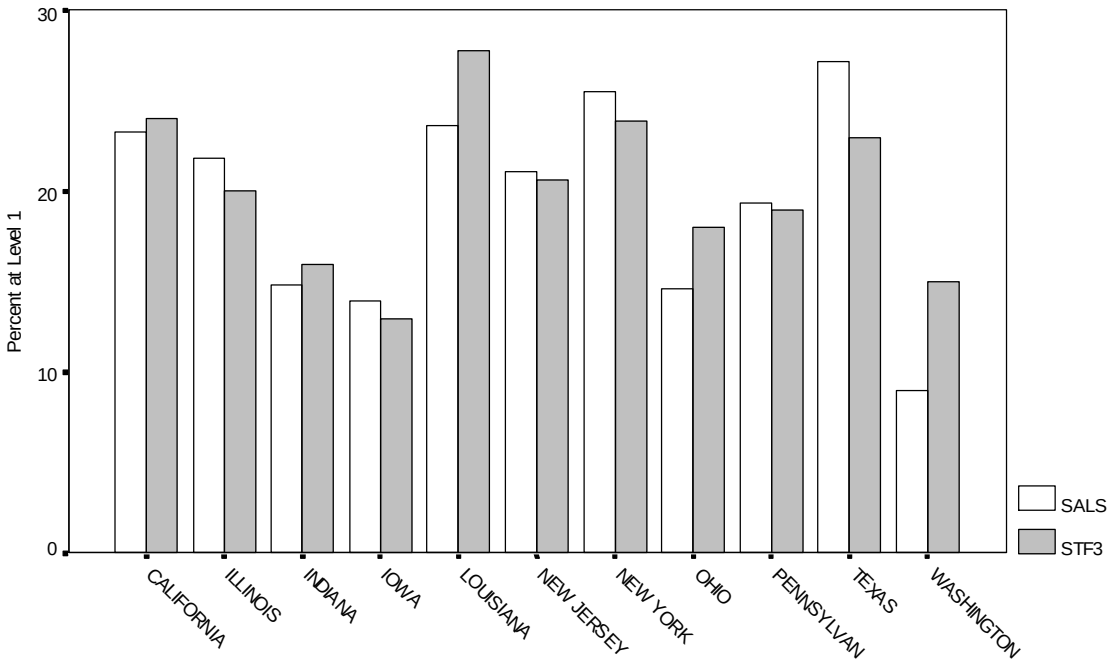

Figure 8. Comparison of synthetic estimates derived from Census STF3 data and State Adult Literacy Survey (SALS) estimates of the statewide percentage of adults with literacy proficiency in Level 1.

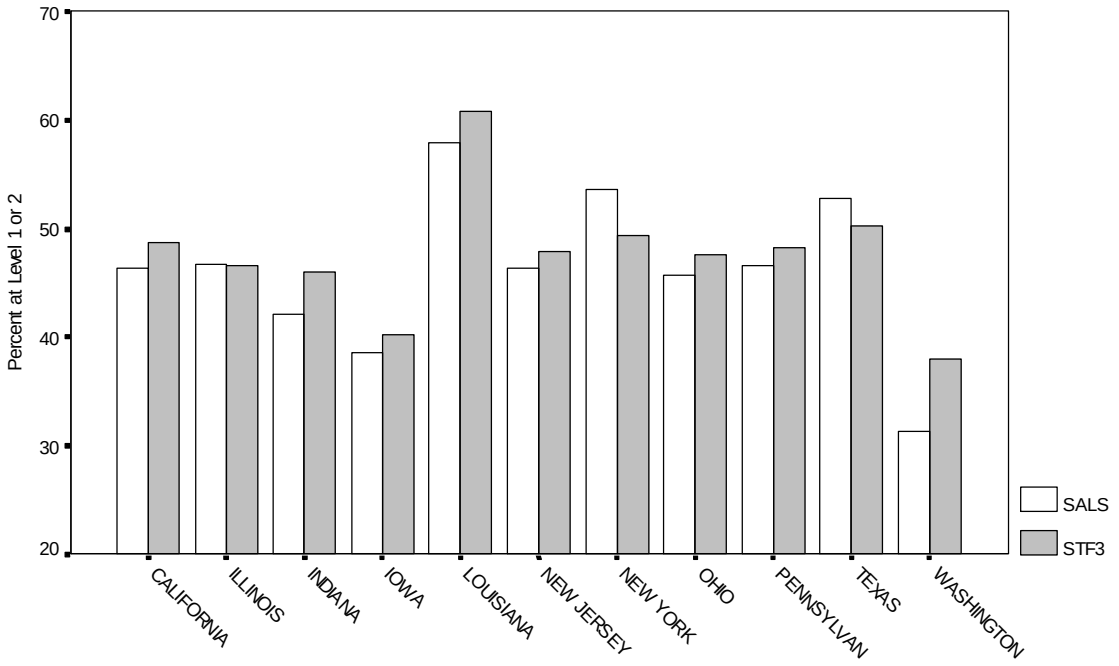

Figure 9. Comparison of synthetic estimates derived from Census STF3 data and State Adult Literacy Survey (SALS) estimates of the statewide percentage of adults with literacy proficiency in Level 1 or 2.

DISCUSSION

There are some important limitations in these synthetic estimates that should be kept in mind when using them. First of all, the regression models were developed from county-level aggregate data within NALS. Synthetic estimates have been developed not only for counties within the U.S., but also for other types of geographical units, including congressional districts, cities, towns and places of 10,000 or more inhabitants, and states. The analysis of the regression model on the county-level aggregates indicated an excellent fit of predictions to observed data. Furthermore, the state-level validation suggests that the model applies reasonably well to much larger units. As promising as these validity studies may be, there is no direct evidence available about the validity of the model's predictions for the congressional district or city/town/place Census areas. Since the NALS database contained no geographical identifiers of levels other than county or state (nor did its sampling design represent these other levels), some caution is appropriate in working with estimates at these levels. While it seems highly plausible that models which predict literacy measures accurately at county and state levels would also perform well at these other levels, the lack of direct validating information should be kept in mind when working with such estimates. On balance, these synthetic estimates should be useful for many purposes in comparing the literacy profiles and needs for service across the various units that may be relevant to decision- and policy-makers in particular contexts. Despite their shortcomings, they may often be the best information available for many geographical areas in which costly local literacy assessment surveys have not been conducted.

REFERENCES

Census of Population and Housing, 1990. (1992). *Summary Tape File 3 on CD-ROM. Technical documentation* / prepared by the Bureau of the Census. Washington: The Bureau [producer and distributor].

Census of Population and Housing, 1990. (August 7, 1992). *Report of the Committee on Adjustment of Postcensal Estimates*. Washington: The Bureau.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics, 19*, 15-18.

Kirsch, I.S., Berlin, M., Mohadjer, L., Rock, D., Yamamoto, K., & others (forthcoming). *Technical report of the 1992 National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics.

Kirsch, I.S., Jungeblut, A., Jenkins, L., & Kolstad. A. (1993). *Adult literacy in America: A first look at the results of the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics.

Murray, T.S. (in press). *Proxy measurement of adult basic skills: Lessons from Canada.* Technical Report, National Center on Adult Literacy.

Murray, T.S., & Shillington, R. (1991). *Estimates of literacy skill for small areas*. Ottawa: Statistics Canada.

National Education Goals Panel (1993a). *The National Education Goals report: Building a nation of learners. Volume One: The national report*. Washington DC: U.S. Government Printing Office.

National Education Goals Panel (1993b). *The National Education Goals report: Building a nation of learners. Volume Two: State reports*. Washington DC: U.S. Government Printing Office.

Reder, S. (1994a). *What does the NALS measure? Issues of dimensionality and construct validity.* Paper presented at the Annual Meeting of the American Educational Research Association. Revised version, "Dimensionality and Construct Validity of the NALS Assessment", to appear in M. C. Smith (Ed.), *Literacy in the 21st Century: Research, policy, practice and the National Adult Literacy Survey.* Greenwood (forthcoming).

Reder, S. (1994b). *Synthetic estimates of NALS literacy proficiencies from 1990 Census microdata*. Portland OR: Northwest Regional Educational Laboratory.

APPENDIX A:
STANDARDIZED REGRESSION COEFFICIENTS

PREDICTOR	MEAN PROFICIENCY	% AT LEVEL 1	% AT LEVEL 1 OR 2
*Educ less than high school			
Educ-some high school	.235		-.155
Educ-high school diploma/GED	.402	-.175	-.326
Educ-some college	.364	-.186	-.313
Educ-2 year college degree	.149		-.156
Educ-4 year college degree	.324		-.255
Educ-graduate school	.472	-.232	-.468
*White			
Black	-.298	.319	.276
Native American			
Asian/Pacific Islander			
Other race			
Work disability	-.067	.133	.096
*No work disability			
*Speaks English very well			
Speaks English well	-.147	.146	.179
Speaks English not well/not at all	-.189	.318	
Recent immigrant	-.085		.104
*Not recent immigrant			
*Did not work previous year			
Worked 1-13 weeks previous year	.090		
Worked 14-26 weeks previous year			
Worked 27-39 weeks previous year			
Worked 40-52 weeks previous year			
*Laborer			
Service			
Sales/administrative support	.055		
Professional/technical/managerial			
*Not in labor force			
Unemployed		-.072	
Employed	.151	-.249	-.186
Northeast			.071
Midwest	.075		
South			.070
*West			

APPENDIX B:
PRECISION OF SYNTHETIC ESTIMATES

	Geographical Unit	State[a]	County[b]	Congressional District	City or Town[c]	County Subdivision[d]
Prediction	Number of entities Predicted	51	2655	436	3154	1370
Average Literacy Proficiency	Median Standard Error of Prediction	1.31	1.67	1.48	1.71	1.70
	Min. Standard Error of Prediction	1.04	0.98	1.00	0.92	1.00
	Max. Standard Error of Prediction	2.24	4.57	3.59	5.71	4.75
	Median Width of 95 % C.I.	7.28	9.30	8.24	9.52	9.43
	Min. Width of 95 % C.I.	5.76	5.43	5.59	5.12	5.58
	Max. Width of 95 % C.I.	12.48	25.43	20.01	31.77	26.42
Proportion at Level 1	Median Standard Error of Prediction	.009	.011	.009	.010	.010
	Min. Standard Error of Prediction	.006	.007	.006	.005	.006
	Max. Standard Error of Prediction	.022	.031	.033	.040	.031
	Median Width of 95 % C.I.	.047	.061	.052	.058	.055
	Min. Width of 95 % C.I.	.036	.036	.036	.030	.035
	Max. Width of 95 % C.I.	.120	.171	.184	.224	.174
Proportion at Level 1 or 2	Median Standard Error of Prediction	.011	.014	.013	.015	.015
	Min. Standard Error of Prediction	.009	.008	.009	.008	.009
	Max. Standard Error of Prediction	.019	.031	.029	.046	.036
	Median Width of 95 % C.I.	.064	.077	.072	.082	.083
	Min. Width of 95 % C.I.	.051	.047	.049	.044	.051
	Max. Width of 95 % C.I.	.106	.171	.160	.259	.202

[a] Includes District of Columbia
[b] Excludes counties with fewer than 5,000 individuals age 16 and above
[c] Excludes entities with fewer than 10,000 total individuals or 5,000 individuals age 16 and above
[d] Excludes county subdivisions with fewer than 5,000 individuals age 16 and above

ENDNOTES

[1]Thanks are due to several people who assisted this effort. Professor Robert Fountain, director of the Statistical Consulting Laboratory at Portland State University, provided very helpful suggestions and discussion regarding the design, implementation and analysis of the estimation models. Chris Wingerd and Charlie Mauck, students of Dr. Fountain, helped with the construction of databases and with the running of statistical programs. The computer software that was developed to display results of these analyses was programmed by David Lowry and Charlie Mauck (Windows version) and by Cavanaugh and Theodore Latiolais (Macintosh version).

[2]This effort was funded by the U.S. Department of Education, Office of Vocational and Adult Education. The opinions, findings and conclusions in this paper and associated database are those of the author; no endorsement should be inferred by the U.S. Department of Education or any other agency.

[3]The NALS sample also included a component which sampled individuals incarcerated in state and federal prisons; only the household component of the NALS is pertinent here, since the prison sample was not designed for state-level disaggregation and is not included in SALS estimates.

[4] These eleven states were California, Illinois, Indiana, Iowa, Louisiana, New Jersey, New York, Ohio, Pennsylvania, Texas and Washington.

[5] Florida also conducted a SALS survey, but after the NALS had been completed. Oregon and Mississippi conducted similar surveys of their adult populations, but limited the age range involved (Oregon surveyed those 16-65, whereas Mississippi surveyed the 16-75 age range).

[6] The public use version of the NALS data set masks county identifiers for those counties having relatively small populations in order to protect respondents' confidentiality. The version of the data set used in these analyses did not mask the identifiers of counties with small populations. In the end, however, this did not matter, since only counties with NALS subsamples of at least 50 survey respondents were used in the regression modeling, and all such counties were sufficiently large to have unmasked identifiers. Thus individuals wishing to replicate or extend this modeling can do so with the public use data set.

[7] The combined literacy proficiency was calculated as the mean of the 15 plausible values imputed for each respondent - 5 for prose, 5 for document and 5 for quantitative literacy. The prose, document and quantitative scales were combined in this fashion because they are very highly intercorrelated and can be well represented by a single proficiency measure (Reder, 1994a). Previous synthetic estimate studies separately estimated the three proficiency scales and found the synthetic estimates to be even more highly intercorrelated (Reder, 1994b).

[8] SPSS for Windows 6.1.3 was used to estimate these models.

[9] All congressional districts, states and the District of Columbia met these screening criteria.

[10] The confidence intervals were calculated for the *individual* Census area rather than for the mean of all areas like it (which would be a smaller or tighter confidence interval). This is often called the *prediction interval.*

[11] A SALS survey conducted by Florida shortly after NALS is not included in this comparison.

[12] The individual SALS estimates, being based on relatively small sample sizes, have standard errors that also must be taken into account in evaluating the fit of the model's predictions to these state-level assessment results.